Overeenstemmingsmaten voor nominale data

Popping, Roelof

7. SAMENVATTING EN EINDCONCLUSIES

In deze studie hebben overeenstemmingsmaten voor nominale data centraal gestaan. In hoofdstuk 2 (pagina 12 en verder) zijn dertien desiderata genoemd waaraan dergelijke maten zouden moeten voldoen. Hieronder worden de desiderata genoemd, samen met een trefwoord. Met behulp van dit trefwoord zijn ze overal steeds snel te herkennen:

1. de maximaal mogelijke bovengrens van de maat is 1, ongeacht aantallen beoordelaars of categorieën;
2. de maat heeft bij onafhankelijkheid tussen beoordelaars de waarde 0 (ONAFH);
3. elke beoordelaar stemt perfect overeen met zichzelf (PERFZELF);
4. perfecte overeenstemming is een transitieve relatie (PERFTRANS);
5. permutaties van categorieën mogen niet tot andere resultaten leiden. Omdat de data op nominaal niveau gemeten zijn, is de volgorde van de categorieën willekeurig en niet van invloed op de uitkomsten (PERMUT);
6. de geschatte waarde van de maat is onafhankelijk van het aantal beoordeelde eenheden (NOBS);
7. gegeven dat er meer dan twee categorieën zijn, moet de mate van overeenstemming berekend kunnen worden over alle categorieën gezamelijk, maar ook per categorie afzonderlijk (CATEG);
8. gegeven dat er meer dan twee beoordelaars zijn, moet de mate van overeenstemming berekend kunnen worden voor alle beoordelaars samen, maar ook per beoordelaar (BEOORD);
9. de maat moet symmetrisch zijn. Een uitzondering kan gemaakt worden voor de situatie waarin één van de beoordelaars geldt als standaard (SYMMETR);
10. de steekproefverdeling of minstens de variantie van de maat moet (exact of desnoods bij benadering) bekend zijn (VAR);
11. de maat moet robuust zijn (ROBUUST);
12. de maat moet eenvoudig en interpreteerbaar zijn (EENVOUD);
13. de maat moet valide zijn (VALIDE).

Tevens zijn een aantal empirische situaties onderscheiden waarin deze maten gebruikt moeten kunnen worden. In deze situaties worden de volgende aspecten onderscheiden:

-liggen de categorieën van te voren vast of niet;
-zijn de beoordelaars per beoordeling wel of niet bekend;
-welke definitie van overeenstemming wordt gehanteerd (paarsgewijze, simultane of meerderheidsovereenstemming);
-is er een standaard;
-worden gewichten op grond van ernst van categorieverschillen toegekend of niet;
-is er sprake van intracategorie vergelijking, de mate van overeenstemming binnen een categorie;

-is er sprake van intercategorie vergelijking, de mate van overeenstemming tussen categorieën;
-is er sprake van intracategorie vergelijking geconditioneerd op een van de beoordelaars;
-het aantal keren dat een eenheid is beoordeeld;
-zijn er ontbrekende waarden;
-hoeveel categorieën zijn er.

In de daaropvolgende hoofdstukken is nagegaan welke maten er zijn, in hoeverre deze aan de desiderata voldoen en of deze toepasbaar zijn binnen de onderscheiden aspecten van empirische situaties. Het belangrijkste empirische onderscheid heeft betrekking op het feit of de beoordelaars de categorieën van te voren kenden waaraan de onderzoekseenheden moeten worden toegewezen. Als de beoordelaars zelf categorieën op hebben moeten stellen, dan is niet te voren gegarandeerd dat de beoordelaars zijn uitgegaan van hetzelfde categorieënstelsel. In deze situatie wordt gesproken over de a priori methode van coderen. Als de beoordelaars allen gebruik maken van hetzelfde, van te voren bekende, categorieënstelsel wordt gesproken over de a posteriori methode van coderen.

In hoofdstuk 3 zijn de maten geanalyseerd die bruikbaar zijn bij de a posteriori methode van coderen. Gebleken is dat van de onderzochte indices de coëfficiënt Cohens kappa (paragraaf 3.2) als enige voldoet aan alle gestelde desiderata, en dat de index gegeneraliseerd kan worden tot alle onderscheiden empirische situaties. Een aantal van deze generalisaties zijn hier voor het eerst behandeld of in formulevorm weergegeven: kappa voor meerderheidsovereenstemming tussen beoordelaars, gewogen kappa per categorie voor meerdere beoordelaars, kappa per categorie voor simultane overeenstemming tussen beoordelaars, kappa voor interklasse vergelijking tussen beoordelaars, kappa voor simultane overeenstemming bij beoordelingen (ook per categorie), kappa voor meerderheidsovereenstemming bij beoordelingen.

De algemene kappa-maat berust op de fractie waargenomen overeenstemming en de fractie overeenstemming die verwacht wordt bij onafhankelijkheid gegeven de marginalen. De exacte definiering van deze twee fracties wordt bepaald door de onderhavige empirische situatie.

De kappa-maat kan zowel gebruikt worden voor beschrijvende doeleinden als voor toetsing (generaliserend van een aselecte steekproef van eenheden naar een populatie).

De meeste overige indices zijn ontwikkeld voor een specifieke situatie; hoewel een groot aantal eventueel wel uitgebreid zou kunnen worden tot andere situaties. Dit heeft evenwel niet zoveel zin, omdat de kappa-index op grond van de desiderata te prefereren is boven deze indices.

In Tabel 3.1 (pagina 86) wordt per overeenstemmingsmaat weergegeven aan welke desiderata hij wel en niet voldoet. Ook als door middel van een eenvoudig te verwezenlijken uitbreiding van de index gerealiseerd kan worden dat de index aan een bepaald desideratum voldoet, is in de tabel aangegeven dat de index voldoet aan dit desideratum.

Een van de belangrijkste verschilpunten tussen de diverse kappa-maten wordt veroorzaakt door de vraag of bekend is welke

beoordelaars welke eenheden hebben geclassificeerd. Indien dit niet zo is, kan geen vergelijking tussen beoordelaars worden gemaakt, maar slechts tussen beoordelingen. Met name de fractie overeenstemming die verwacht wordt bij onafhankelijkheid moet nu op een andere manier worden gedefinieerd.

Bij bepaalde extreem scheve marginale verdelingen is het mogelijk dat de uitkomst die met een kappa-maat gevonden wordt erg laag is terwijl de fractie waargenomen overeenstemming toch erg hoog is. Dit wordt veroorzaakt door de correctie voor bij onafhankelijkheid verwachte overeenstemming. Sommige onderzoekers vinden dit een bezwaar van kappa en verwerpen de maat daarom. Deze onderzoekers zijn niet bereid de gevolgen van bovengenoemde correctie te accepteren. Zij gebruiken daarom indices waarin deze correctie niet voorkomt; vertekende uitkomsten die juist door deze correctie worden tegengegaan, nemen ze dan voor lief. Het zal duidelijk zijn dat hun bezwaren niet worden gedeeld.

In hoofdstuk 4 zijn de indices besproken die gebruikt worden als de a priori methode van coderen is gevolgd. Hier kan alleen maar sprake zijn van beoordelaars, er moet precies bekend zijn welke beoordelaar welke eenheden heeft geclassificeerd en welk categorieënstelsel daarbij is opgesteld. Van de in dit hoofdstuk behandelde indices blijkt dat alleen de D2-index (paragraaf 4.2.4) voldoet aan de gestelde desiderata. In Tabel 4.4 (pagina 113) is aangegeven welke indices aan welke desiderata voldoen.

De D2-index is al eenmaal elders gepresenteerd (Popping, 1983a). De generalisaties naar andere situaties en de variantie van de index zijn voor het eerst in dit boek gepresenteerd. Bij lage uitkomsten is interpretatie van de index moeilijk, omdat niet duidelijk wordt waardoor verschillen zijn veroorzaakt: gebruik van verschillende categorieënstelsels en/of verschillende toewijzingen (dit geldt trouwens voor alle in hoofdstuk 4 behandelde maten).

Ook voor D2 kunnen dezelfde opmerkingen worden gemaakt bij extreem scheve marginalen als hiervoor bij kappa.

De D2-index is vooral geschikt voor vooronderzoekingen waar het doel is de opbouw van adequate categorieënstelsels. Nadat de beoordelaars hun taak hebben uitgevoerd, is het aan te bevelen een stapsgewijze procedure te starten waarin de beoordelaars door onderlinge vergelijking en discussie moeten komen tot een categorieënstelsel dat in het hoofdonderzoek zal worden gebruikt.

Als de beoordelaars zelf categorieënstelsels op moeten stellen, kunnen deze onderling nogal verschillen. In hoofdstuk 5 is onderzocht of er een aantal factoren zijn welke er toe zullen bijdragen dat er een hogere overeenstemming ontstaat. Nagegaan is of het invoeren van een aantal randvoorwaarden (al dan niet in een bepaalde combinatie) zal leiden tot hogere overeenstemming tussen de beoordelaars. Deze randvoorwaarden zijn:

- het stellen van een bovengrens met betrekking tot het aantal categorieën dat gebruikt mag worden;
- het geven van informatie over het onderzoek waar de vraag uit afkomstig is en de betekenis van de vraag daarin, alsmede het tijdstip waarop deze informatie wordt gegeven; en
- het geven van het categorieënstelsel dat is opgesteld door een beoordelaar die de codeerwerkzaamheden eerder onder dezelfde

condities heeft verricht.
Dit is onderzocht voor vragen die betrekking hebben op houding, gedrag en feitelijke informatie.

Het blijkt dat de formulering van de vraag het meest essentieel is: geeft deze aanleiding tot een voor de hand liggend categorieënstelsel, of is dit niet zo? Als dit zo is, worden sterk gelijkende categorieën opgesteld, als het niet zo is worden heel verschillende categorieënstelsels opgesteld. Ook het aantal categorieën dat men heeft mogen gebruiken blijkt van belang te zijn, vooral in interactie met de vraag. Er is evenwel geen regelmatig patroon in de gevonden verschillen aan te wijzen. Informatie die de beoordelaars hebben gehad, draagt niet aantoonbaar bij tot een hogere overeenstemming. In het onderzoek hebben de beoordelaars voorzover ze informatie hebben gekregen, deze gehad vlak voordat ze met hun codeerwerkzaamheden begonnen.

Niet is onderzocht hoe de overeenstemming zal zijn tussen personen die volledig zijn geinvolveerd in een bepaald onderzoek, die dus beter weten wat het doel van het onderzoek is en waarom de vraag is opgenomen. Ik verwacht dat de overeenstemming hier aanzienlijk hoger zal zijn.

Wel is in het experiment een vergelijking gemaakt met beoordelaars die gebruik hebben gemaakt van de a posteriori methode van coderen (dus gebonden waren aan vaste categorieën). Het blijkt dat de D2-waarden in deze situatie nauwelijks hoger zijn, zodat de lage D2-waarden vooral te wijten zijn zijn aan verschillen in toekenningen.

De belangrijkste conclusie uit het experiment is dat verschillen in toewijzingen grotere invloed hebben op de mate van overeenstemming tussen de beoordelaars dan verschillen in categorieënstelsels.

Als andere belangrijke conclusie van dit experiment geldt dat factoren die bijdragen aan een hogere overeenstemming tussen de beoordelaars sterk afhankelijk zijn van de (formulering van) de vraag. In mindere mate is ook het aantal categorieën dat men heeft mogen gebruiken van invloed. Men mag geen al te hoge overeenstemming verwachten als de vraag complex is en er een uitvoerig antwoord is waar de relevante informatie nog uitgelicht moet worden. Bij dergelijke vragen is het aan te bevelen de beoordelaars de opdracht te geven zich te richten op een bepaald aspect in de vraag. Na discussie over het te hanteren categorieënstelsel en hernieuwde toewijzing van de antwoorden, mag een hogere overeenstemming verwacht worden.

In hoofdstuk 6 is een taxonomie gepresenteerd waarin alle empirische situaties zijn onderscheiden die betekenis hebben. De volgende ingangen zijn gebruikt:

--- twee beoordelaars worden met elkaar vergeleken;
- meer dan twee beoordelaars worden paarsgewijs met elkaar vergeleken;
- meer dan twee beoordelaars worden simultaan met elkaar vergeleken;
-- er is geen standaard;
- er is een standaard;
- er wordt geconditioneerd op de categorieën van de standaard;

--- beoordelingen worden vergeleken, de observaties zijn twee keer beoordeeld;
- beoordelingen worden vergeleken, de observaties zijn meer dan twee keer beoordeeld;
- beoordelingen worden simultaan vergeleken, de observaties zijn meer dan twee keer beoordeeld;
-- alle observaties zijn een zelfde aantal keren beoordeeld;
- de observaties zijn niet allemaal hetzelfde aantal keren beoordeeld;
--- er is geen intra- of interklasse vergelijking;
- er is intraklasse vergelijking;
- er is interklasse vergelijking;
--- het aantal categorieën bedraagt twee;
- het aantal categorieën is groter dan twee, er zijn geen gewichten;
- het aantal categorieën is groter dan twee, er zijn wel gewichten;
- het aantal categorieën lag nog niet vast aan het begin van de codeertaak, er worden geen gewichten gebruikt;
- het aantal categorieën lag nog niet vast aan het begin van de codeertaak, achteraf zijn gewichten toegevoegd.

Daarnaast zijn nog een aantal speciale situaties behandeld. Voor alle combinaties is aangegeven in welke publicaties daar aandacht aan is besteed. Deze taxonomie dient in de eerste plaats als hulpmiddel voor hen die zich verder willen verdiepen in overeenstemmingsmaten voor nominale data.